On the efficient computation of the projector over the Birkhoff polytope

Defeng Sun

Department of Applied Mathematics

The Hong Kong Polytechnic University
香港理工大學

The International Symposium on Mathematical Programming 2018
Bordeaux, July 1-6, 2018

Based on joint work with
Xudong Li (Princeton), and Kim-Chuan Toh (NUS)

Optimization over Birkhoff polytope

Consider optimization problem:

$$\min\{f(X) \mid X \in \mathfrak{B}_n\}$$

Birkhoff polytope:

$$\mathfrak{B}_n := \{X \in \Re^{n\times n} \mid Xe = e,\ X^T e = e, X \geq 0\}$$

$e \in \Re^n$: the vector of all ones

$f : \Re^{n\times n} \to (-\infty, +\infty]$, possibly nonconvex, nonsmooth

Permutation matrices:

$$\mathfrak{D}_n := \{X \in \Re^{n\times n} \mid Xe = e,\ X^T e = e, X_{ij} \in \{0, 1\}\}$$

Convex hull: [Birkhoff, 1946 & J. von Neumann, 1953]

$$\mathfrak{B}_n = \mathrm{conv}(\mathfrak{D}_\mathrm{n})$$

Examples

(i) Graph based Clustering, given affinity matrix K
[Zass et al., NIPS, 2006]:

$$\min\left\{\|X-K\|_F^2 \mid X \in \mathfrak{B}_n\right\}$$

[Wang et al., KDD, 2016]: given $k \geq 0$,

$$\min\left\{\|X-K\|_F^2 + \gamma\|I-X\|_* + r\|X\|_F^2 \mid X \in \mathfrak{B}_n,\ \mathrm{Tr}(X)=k\right\}$$

(ii) Graph Matching, adjacency matrices A, B, [Fiori et al., NIPS, 2013]:

$$\min\left\{\|AX-XB\|_F^2 \mid X \in \mathfrak{B}_n\right\}$$

$$\min\left\{\sum_{i,j}\|((AX)_{ij},(XB)_{ij})\|_2 \mid X \in \mathfrak{B}_n\right\}$$

Examples

(iii) Relaxation of the seriation problem, 2-SUM minimization, given symmetric, binary matrix A, [Fogel et al., NIPS, 2013], [Lim & Wright, 2014]:

$$\min\{\langle Xg,\, L_A Xg\rangle \mid X \in \mathfrak{B}_n\}$$

$g = (1, \ldots, n)^T$ and $L_A = \mathrm{diag}(Ae) - A$

Relaxations of the quadratic assignment problems (QAP)...

Difficulties: constrained minimization, n^2 unknown variables in X, large-scale even when $n = 10^3$

Projection onto $\mathfrak{B}_n$

Metric projector over $\mathfrak{B}_n$: given $G \in \Re^{n\times n}$

$$\Pi_{\mathfrak{B}_n}(G) := \arg\min \left\{ \frac{1}{2}\|X - G\|_F^2 \mid X \in \mathfrak{B}_n \right\}$$

Easy computation of $\Pi_{\mathfrak{B}_n}(G)$ "$\Longrightarrow$" first order methods for

$$\min\{f(X) + \delta_{\mathfrak{B}_n}(X)\}$$

$\delta_{\mathfrak{B}_n}(\cdot)$: indicator function over $\mathfrak{B}_n$

PG, APG (FISTA), ADMM, splitting methods ...

Generalized Jacobian $\partial\Pi_{\mathfrak{B}_n}(G)$ "$\Longrightarrow$" second order methods

(sparse) semismooth (generalized) Newton methods

Computation of $\Pi_{\mathfrak{B}_n}(G)$

Primal formulation:

$$\Pi_{\mathfrak{B}_n}(G) = \arg\min \left\{ \frac{1}{2}\|X - G\|_F^2 \ \mid \ \mathcal{B}X = b, \ X \in C \right\}$$

$\mathcal{B}(X) = [Xe; \ X^T e]$, $C := \{X \in \Re^{n\times n} \mid X \geq 0\}$, $b = [e; e]$

Dual problem, convex differentiable function φ:

$$\min \left\{ \varphi(y) := \frac{1}{2}\|\Pi_C(\mathcal{B}^* y + G)\|_F^2 - \langle b, \, y\rangle - \frac{1}{2}\|G\|_F^2 \ \mid \ y \in \mathrm{Range}(\mathcal{B}) \right\}$$

$\Pi_{\mathfrak{B}_n}(G) = \Pi_C(\mathcal{B}^* \bar{y} + G)$ with $\bar{y} \in \arg\min_{y\in \mathrm{Range}(\mathcal{B})} \varphi(y)$

Simple convex QP, state-of-the-art commercial solvers (e.g., Gurobi, Mosek)

(accelerated) gradient methods, semismooth Newton methods

Newton's method

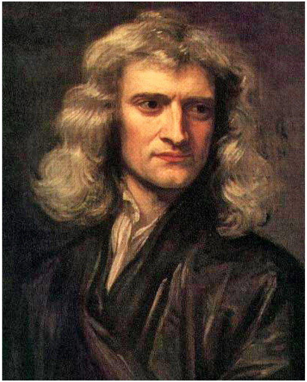


Figure: Sir Isaac Newton (Niu Dun) (4 January 1643 - 31 March 1727)

Which Newton's method?

Figure: Snail (Niu)

Figure: Longhorn beetle (Niu)

Figure: Charging Bull (Niu)

Figure: Yak (Niu)

Algorithm SSNCG

Recall $\nabla\varphi(y) = \mathcal{B}\Pi_C(\mathcal{B}^*y + G) - b, \quad \forall\, y \in \mathrm{Range}(\mathcal{B})$

SSNCG for nonsmooth equation:

$$\nabla\varphi(y) = 0, \quad y \in \mathrm{Range}(\mathcal{B})$$

j-th iter. Choose $\mathcal{V}_j \in \mathcal{B}\partial\Pi_C(\mathcal{B}^*y_j + G)\mathcal{B}^*$, solve linear system (CG)

$$(\mathcal{V}_j + \varepsilon_j I)d + \nabla\varphi(y^j) = 0, \quad d \in \mathrm{Range}(\mathcal{B})$$

$\varepsilon_j > 0$: small perturbation converging to zero.

Global convergence: Line search (using $\varphi(y)$)

Local convergence: at least superlinear if

$$\mathcal{B}\,\mathrm{lin}(\mathcal{T}_C(\Pi_C(\mathcal{B}^*\bar{y} + G)) = \mathrm{Range}(\mathcal{B})$$

SSNCG fast and robust

Numerical comparisons

Relative KKT residual: $\eta = \max\{\eta_P, \eta_C\}$,

$$\eta_P = \frac{\|\mathcal{B}X - b\|}{1 + \|b\|}, \quad \eta_C = \frac{\|X - \Pi_C(\mathcal{B}^* y + G)\|}{1 + \|X\|}$$

Algorithms: our semismooth Newton CG method (Ssncg), accelerated gradient, Gurobi

Test instances G:

1. `G = randn(n)`
2. Similarity matrices derived from LIBSVM datasets: **gisette**, **mushrooms**, **a6a**, **a7a**, **rcv1** and **a8a**

   normalize (unit l_2-norm), Gaussian kernel

$$G_{ij} = \exp\left(-\|x_i - x_j\|^2\right), \quad \forall\, 1 \le i, j \le n$$

Gurobi: default parameters

Ssncg and APG: terminate $\eta < \mathrm{tol}$ or the time exceeding 3 hours

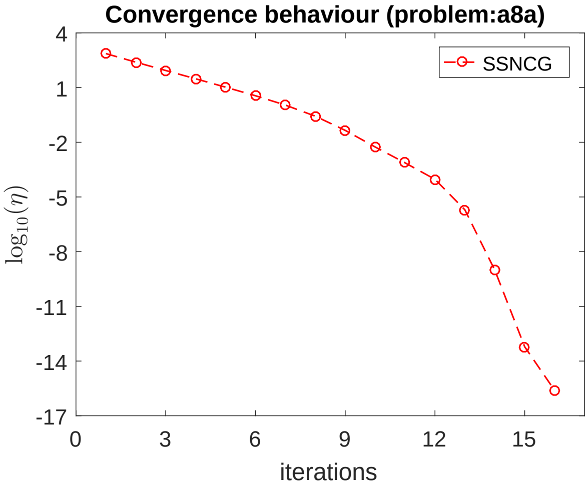
Numerical results

"a": Accelerated gradient ($\mathrm{tol} = 10^{-9}$), "b": Gurobi,

"c": SSNCG ($\mathrm{tol} = 10^{-15}$)

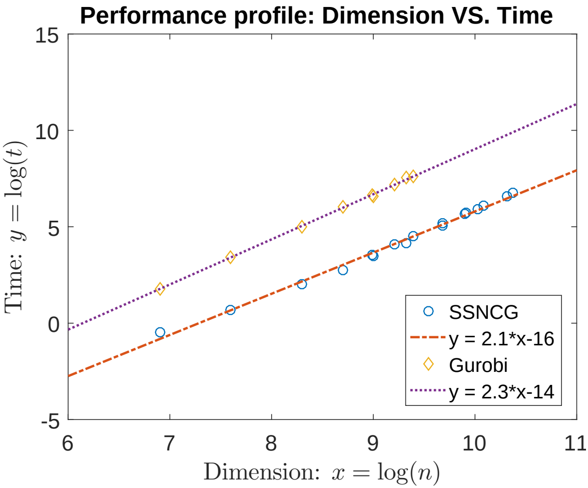
"*": Gurobi out of memory (128 G RAM)

		η	time
problem	n	a | b(η_P) | c	a | b | c
rand5	10000	**3.1-7** | **6.2-6** (3.4-15) | 4.5-16	3:00:00 | 21:27 | 58
rand6	12000	**2.6-4** | **4.6-6** (3.8-15) | 4.6-16	3:00:00 | 33:31 | 1:33
rand9	24000	**1.9-2** | * | 4.9-16	3:00:04 | * | 7:15
rand10	30000	**4.9-2** | * | 5.9-16	3:00:14 | * | 12:01
rand11	32000	**6.3-2** | * | 4.8-16	3:00:17 | * | 14:10
gisette	6000	9.8-10 | **3.3-6** (2.5-15) | 6.5-16	7:19 | 6:58 | 16
mushrooms	8124	9.8-10 | **9.5-5** (4.8-15) | 1.9-16	11:07 | 11:58 | 32
a6a	11220	9.9-10 | **4.7-6** (4.0-15) | 3.8-16	34:29 | 31:21 | 1:03
a7a	16100	9.9-10 | * | 2.9-16	1:28:14 | * | 2:34
rcv1	20242	**1.3-6** | * | 1.9-16	3:00:03 | * | 5:02
a8a	22696	**2.7-4** | * | 2.5-16	3:00:03 | * | 6:15



Superlinear convergence of SSNCG

Performance profile



Computational complexities:

$$t_{\text{Ssncg}} = \mathcal{O}(n^{2.1}), \quad t_{\text{Gurobi}} = \mathcal{O}(n^{2.3}), \quad \frac{t_{\text{Gurobi}}}{t_{\text{Ssncg}}} = 7n^{0.2}$$

SsnCG VS. PPROJ

PPROJ: C codes, extremely fast implementation, algorithm for projection onto general polyhedral sets [Hager & Zhang, SIOPT, 2016]

"a": PPROJ, "b": SsnCG ($\mathrm{tol} = 10^{-15}$)

Executed on HPC clusters in NUS

		η	time
problem	n	a | b	a | b
rand5	10000	5.8-12 | 4.5-16	3:45 | 39
rand6	12000	1.6-12 | 4.7-16	5:31 | 59
rand7	16000	5.8-13 | 5.9-16	19:34 | 1:20
rand8	20000	4.3-13 | 9.5-16	34:01 | 2:10
gisette	6000	1.4-14 | 6.5-16	2:39 | 11
mushrooms	8124	**6.9-7** | 2.0-16	16:22:46 | 21
a6a	11220	3.3-14 | 3.8-16	14:32 | 39
a7a	16100	3.7-14 | 2.9-16	43:53 | 1:21
rcv1	20242	1.3-13 | 1.9-16	2:01:32 | 2:19

Generalized Jacobian of $\Pi_D(\cdot)$

$\partial \Pi_{\mathfrak{B}_n}(G)$: second order algorithms for optimization over $\mathfrak{B}_n$

Polyhedral $D := \{x \in \Re^n \mid Ax \geq b,\ Bx = d\}$, B full row rank

For any $x \in \Re^n$, KKT condition for $\Pi_D(x)$

$$\begin{cases} \Pi_D(x) - x + A^T\lambda + B^T\mu = 0, \\ A\Pi_D(x) - b \geq 0, \quad B\Pi_D(x) - d = 0, \\ \lambda \leq 0, \quad \lambda^T(A\Pi_D(x) - b) = 0 \end{cases} \tag{1}$$

Multipliers $M(x) := \{(\lambda, \mu) \in \Re^m \times \Re^p \mid (x, \lambda, \mu) \text{ satisfies } (1)\}$

Active index $I(x) := \{i \mid A_i\Pi_D(x) = b_i,\ i = 1, \ldots, m\}$

$\Pi_D(\cdot)$: non-smooth, piecewise-affine, difficult to find elements in $\partial \Pi_D(x)$

Any computable approach?

Generalized Jacobian of $\Pi_D(\cdot)$

A_K: rows of A, indexed by K

$$\mathcal{D}(x) := \{\ K \subseteq [1:m] \mid \exists\, (\lambda,\mu) \in M(x) \text{ s.t. } \operatorname{supp}(\lambda) \subseteq K \subseteq I(x), \\ [A_K^T\ \ B^T] \text{ is of full column rank}\}$$

$$\mathcal{P}(x) := \left\{ \begin{array}{l} P \in \Re^{n\times n} \mid \\ P = I_n - [A_K^T\ \ B^T]\left(\begin{bmatrix} A_K \\ B \end{bmatrix} [A_K^T\ \ B^T]\right)^{-1} \begin{bmatrix} A_K \\ B \end{bmatrix},\ K \in \mathcal{D}(x) \end{array} \right\}$$

$\mathcal{P}(\cdot)$: Upper semicontinuity, [Han & Sun, JOTA, 1997]

Semismooth Newton method: find an element in $\mathcal{P}(x)$

Difficulties:

1. K corresponds to λ, λ unavailable
2. full column rank of $[A_K^T\ \ B^T]$

HS-Jacobian of Π_D

Theorem 1

For any given $x \in \Re^n$, denote

$$P_{HS} := I_n - [A_{I(x)}^T \;\; B^T] \left(\begin{bmatrix} A_{I(x)} \\ B \end{bmatrix} [A_{I(x)}^T \;\; B^T] \right)^{\dagger} \begin{bmatrix} A_{I(x)} \\ B \end{bmatrix}.$$

Then, $P_{HS} \in \mathcal{P}(x)$.

P_{HS}: referred to as HS-Jacobian (HS: Han & Sun)

Advantages of P_{HS}:

1. unrelated to index sets K and multipliers λ
2. full column rank assumption removed

HS-Jacobian of $\Pi_{\mathfrak{B}_n}$

Given $G \in \Re^{n\times n}$, denote $\overline{G} := \Pi_{\mathfrak{B}_n}(G)$

Linear operator $\Xi : \Re^{n\times n} \to \Re^{n\times n}$

$$\Xi(H) := \Theta \circ H, \quad H \in \Re^{n\times n}, \quad \Theta_{ij} = \begin{cases} 0, & \text{if } \overline{G}_{ij} = 0, \\ 1, & \text{otherwise.} \end{cases}$$

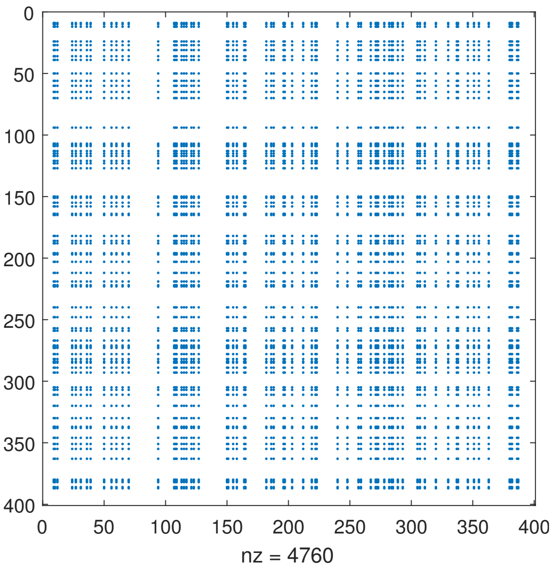
Proposition 1

The linear operator $P_{HS} : \Re^{n\times n} \to \Re^{n\times n}$ given by

$$P_{HS}(H) := \Xi(H) - \Xi\mathcal{B}^*(\mathcal{B}\Xi\mathcal{B}^*)^\dagger \mathcal{B}\Xi(H), \quad \forall\, H \in \Re^{n\times n},$$

is the HS-Jocobian of $\Pi_{\mathfrak{B}_n}$ at G.

Sparsity of HS-Jacobian of $\Pi_{\mathfrak{B}_n}$



$$n = 20, \quad P_{HS} \in \Re^{n^2 \times n^2}, \quad \text{sparsity} = 0.0297$$

QP with Birkhoff polytope constraints

Convex QP:

$$(\mathbf{P}) \quad \min \left\{ \frac{1}{2}\langle X,\, \mathcal{Q}X\rangle + \langle G,\, X\rangle \mid X \in \mathfrak{B}_n \right\},$$

Self-adjoint linear operator $\mathcal{Q} \succeq 0$

$$(\mathbf{D}) \min \left\{ \delta^*_{\mathfrak{B}_n}(Z) + \frac{1}{2}\langle W,\, \mathcal{Q}W\rangle \mid Z + \mathcal{Q}W + G = 0,\, W \in \mathrm{Range}(\mathcal{Q}) \right\}$$

$\delta^*_{\mathfrak{B}_n}$: the conjugate of the indicator function $\delta_{\mathfrak{B}_n}$

ALM function for **(D)**, given $\sigma > 0$

$$\mathcal{L}_\sigma(Z, W; X) = \delta^*_{\mathfrak{B}_n}(Z) + \frac{1}{2}\langle W,\, \mathcal{Q}W\rangle - \langle X,\, Z + \mathcal{Q}W + G\rangle + \frac{\sigma}{2}\|Z + \mathcal{Q}W + G\|^2$$

ALM for (D)

Algorithm ALM: An augmented Lagrangian method for (D).

Given $\sigma_0 > 0$, iterates $k = 0, 1, \ldots$

Step 1. Compute

$$(Z^{k+1}, W^{k+1}) \approx \operatorname{argmin} \left\{ \begin{array}{l} \Psi_k(Z, W) := \mathcal{L}_{\sigma_k}(Z, W; X^k) \\ \mid (Z, W) \in \Re^{n\times n} \times \text{Range}(\mathcal{Q}) \end{array} \right\}.$$

Step 2. Compute

$$X^{k+1} = X^k - \sigma_k(Z^{k+1} + \mathcal{Q}W^{k+1} + G).$$

Update $\sigma_{k+1} \uparrow \sigma_\infty \leq \infty$.

Convex piecewise linear-quadratic minimization:

error bound holds $\Longrightarrow$ ALM converges asymptotically superlinearly

Semismooth Newton method CG for inner problem

For any $W \in \mathrm{Range}(\mathcal{Q})$,

$$\psi(W) := \inf_{Z} \mathcal{L}_\sigma(Z, W; \widehat{X}), \quad Z(W) := \widehat{X} - \sigma(\mathcal{Q}W + G)$$

Subproblem solution $(\overline{Z}, \overline{W})$:

$$\overline{W} = \arg\min \left\{ \psi(W) \ | \ W \in \mathrm{Range}(\mathcal{Q}) \right\},$$
$$\overline{Z} = \sigma^{-1}\big(Z(\overline{W}) - \Pi_{\mathfrak{B}_n}(Z(\overline{W}))\big)$$

For all $W \in \mathrm{Range}(\mathcal{Q})$,

$$\nabla\psi(W) = \mathcal{Q}W - \mathcal{Q}\Pi_{\mathfrak{B}_n}(Z(W))$$

Semismooth Newton CG solve nonsmooth piecewise affine equation

$$\nabla\psi(W) = 0, \quad W \in \mathrm{Range}(\mathcal{Q}).$$

Semismooth Newton method CG for inner problem

Given $\widehat{W}$, linear operator $\mathcal{M} : \Re^{n\times n} \to \Re^{n\times n}$

$$\mathcal{M}(\Delta W) := (\mathcal{Q} + \sigma \mathcal{Q} P_{HS} \mathcal{Q})\Delta W, \quad \forall \Delta W \in \Re^{n\times n}$$

P_{HS}: the HS-Jacobian of $\Pi_{\mathfrak{B}_n}$ at $Z(\widehat{W})$

j-th iter., solve linear system (CG)

$$\mathcal{M}_j dW + \nabla \psi(W^j) = 0, \quad dW \in \text{Range}(\mathcal{Q})$$

Global convergence: Line search (using $\psi(W)$)

Local convergence:
positive definiteness of $\mathcal{M}$ on $\text{Range}(\mathcal{Q})$ $\Longrightarrow$ at least superlinear

Numerical results

Given $A, B \in \mathcal{S}^n$, QAP:

$$\min\{\langle X,\ AXB\rangle \mid X \in \{0,1\}^{n\times n} \cap \mathfrak{B}_n\}$$

Convex relaxation [Anstreicher et al. MP, 2001]:

$$\min\{\langle X,\ \mathcal{Q}X\rangle \mid X \in \mathfrak{B}_n\}$$

Self-adjoint linear operator $\mathcal{Q}(X) := AXB - SX - XT, \quad \mathcal{Q} \succeq 0$

Matrices $S, T \in \mathcal{S}^n$ obtained from [Anstreicher et al. MP, 2001]

Relative KKT residual:

$$\eta = \frac{\|X - \Pi_{\mathfrak{B}_n}(X - \mathcal{Q}X)\|}{1 + \|X\| + \|\mathcal{Q}X\|}$$

Matrices A, B from QAPLIB

Numerical results for QAP

"a": Gurobi, "b": ALM

		iter	η	time
problem	n	a | b (itersub)	a|b	a|b
lipa80a	80	11 | 25 (68)	1.3-6 | 7.3-8	2:46 | 01
lipa90a	90	11 | 20 (54)	2.7-6 | 8.8-8	5:32 | 01
sko100a	100	14 | 26 (95)	8.5-6 | 8.5-8	2:06 | 11
tai100a	100	11 | 18 (52)	1.3-6 | 9.5-8	10:31 | 02
tai100b	100	11 | 27 (98)	1.3-6 | 9.1-8	10:31 | 13
tai80b	80	11 | 27 (98)	1.2-6 | 8.5-8	2:36 | 07
tai256c	256	* | 2 (4)	* | 2.1-16	* | 00
tai150b	150	19 | 27 (94)	4.3-7 | 9.3-8	2:46:17 | 13
tho150	150	16 | 24 (96)	5.6-6 | 9.9-8	18:52 | 22

"*": Gurobi out of memory (128 G RAM)

"tai150b": Gurobi reports error, "small positive term" needed

References

[Defeng Sun, Houdou Qi], A quadratically convergent Newton method for computing the nearest correlation matrix, SIMAX, 2006

[Xudong Li, Defeng Sun, Kim-Chuan Toh], On the efficient computation of a generalized Jacobian of the projector over the Birkhoff polytope, arXiv:1702.05934, 2017

Thank you for your attention!